설계 프로젝트 결과 보고

학번: 2015112113

이름: 정용헌

제출일: 2020.06.22.

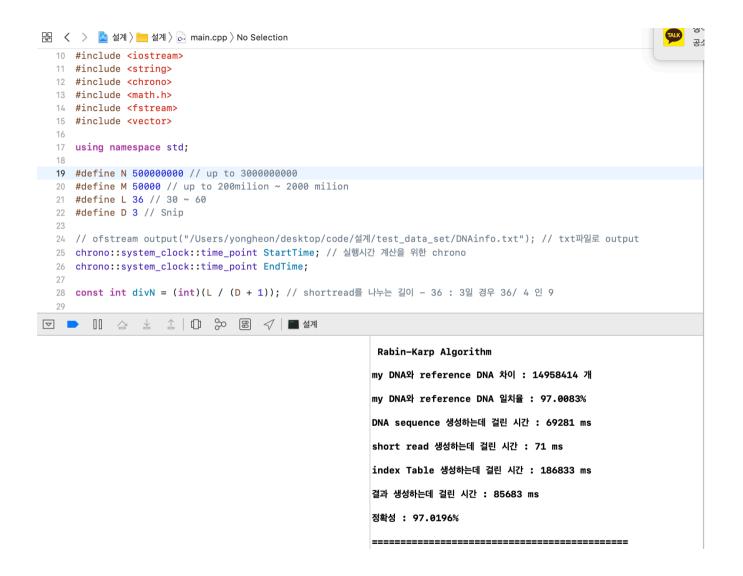
DEFAULT

```
10 #include <iostream>
  11 #include <string>
  12 #include <chrono>
  13 #include <math.h>
  14 #include <fstream>
  15 #include <vector>
  17 using namespace std;
19 #define N 250000 // up to 3000000000
  20 #define M 50000 // up to 200milion ~ 2000 milion
  21 #define L 30 // 30 ~ 60
  22 #define D 3 // Snip
  24 // ofstream output("/Users/yongheon/desktop/code/설계/test_data_set/DNAinfo.txt"); // txt파일로 output
  25 chrono::system_clock::time_point StartTime; // 실행시간 계산을 위한 chrono
 26 chrono::system_clock::time_point EndTime;
 28 const int divN = (int)(L / (D + 1)); // shortread를 나누는 길이 - 36 : 3일 경우 36/ 4 인 9
☑ ▶ [] △ ⊻ △ | □ ‰ 圖 ✓ | ■ 설계
                                                     ===== 2015112113 정용헌 설계 프로젝트 ======
                                                     Rabin-Karp Algorithm
                                                    my DNA와 reference DNA 차이 : 7934 개
                                                    my DNA와 reference DNA 일치율 : 96.8264%
                                                    DNA sequence 생성하는데 걸린 시간 : 36 ms
                                                    short read 생성하는데 걸린 시간 : 49 ms
                                                    index Table 생성하는데 걸린 시간 : 89 ms
                                                    결과 생성하는데 걸린 시간 : 417 ms
                                                    정확성 : 99.88%
                                                    _____
```

프로그램의 기본적 개요는 다음과 같습니다. Define으로 변수 설정 (염기서열의 갯수, 쇼트리드의 길이, 쇼트리드 갯수, SNP) 하여 난수를 생성하여 My DNA와 Reference DNA를 생성하여 둘의 일치율을 비교합니다. 그 후 ShortRead와 Index Table을 생성하여 복원을 진행하고 결과를 생성한 후 정확도를 나타냅니다. 이 때 Index Table을 생성할 때 쇼트리드를 나누어주는데, L / D+1로 설정하였습니다. 이 때 조교님께서 L값이 커지게 되면 메모리 누수에 대한 지적을 해주셨는데, 그럴 땐 D값을 조정하거나 나눔값을 정수로 고정시키면 해결할 수 있으나 제가 여러 변수로 실행해본결과 오류가 발생한 적은 없었습니다.

예를 들어 36 / 3일때 9의 길이의 쇼트리드 devide를 생성합니다. AAAAAAAA = 0, TTTTTTTT = 4^9가 됩니다. 저희 팀에서 디폴트 값으로 설정햇던 250000 / 50000 / 30 / 3 으로 진행했을 때의 결과입니다.

N = 5억



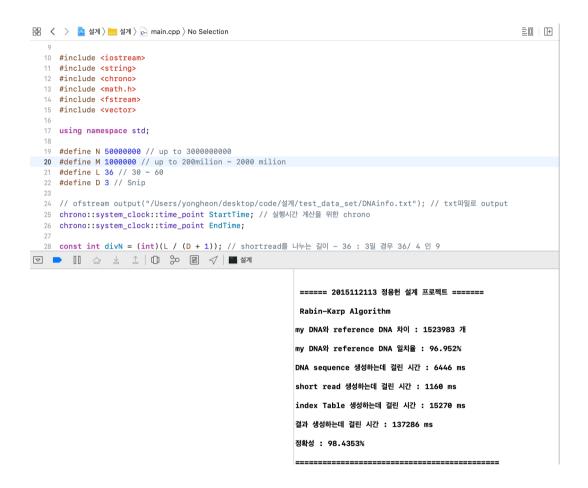
다른 알고리즘에 비해 실행시간이 빠르고 정확도도 높게 나타납니다. N을 5억으로 두고 진행하였을때에도 무리없이 진행되었다.

N = 10억

```
맮 〈 〉 🤷 설계 〉 🛅 설계 〉 ⊙ main.cpp 〉 No Selection
         COUL .. /IITHOCK LODIC OO I
 180
          getTime();
 181
 182
 183
       void makeResult()
 184
 185
          StartTime = chrono::system_clock::now();
 186
          int index = 0;
 187
          for (int i = 0; i < M; i++)
 188
 189
             index = findshortIndex(i);
 190
             if (index !=-1)
 191
 192
                for (int j = 0; j < L; j++)
 193
                   referSequence[index++] = shortRead[i][j];
 195
 196
             }
 197
 198
          EndTime = chrono::svstem clock::now():
☑ ▶ [] △ ↓ ↑ □ ‰ 圖 ✓ ■ 설계
                                                      ===== 2015112113 정용헌 설계 프로젝트 ======
                                                      Rabin-Karp Algorithm
                                                     my DNA와 reference DNA 차이 : 29574866 개
                                                     my DNA와 reference DNA 일치율 : 97.0425%
                                                     DNA sequence 생성하는데 걸린 시간 : 137233 ms
                                                     short read 생성하는데 걸린 시간 : 53 ms
                                                     index Table 생성하는데 걸린 시간 : 210005 ms
```

N이 5억일 때는 모든 과정이 3분내로 진행되었는데 10억으로 변경하자 1시간이 넘도록 결과가 생성되지 않았습니다. txt파일 생성하고 읽어오는 과정에서 메모리가 너무 커져서 발생하는 문제인 것 같다.

N = 5천만 / M = 1백만



빠른 결과 도출을 위해 N의 수를 줄였다. 1백만까진 무리없이 진행된다. 2백만으로 올리자 한시간이 넘도록 결과가 생성되지 않는다. 쇼트리드의 모든 인덱스를 해쉬로 비교하는 방식이라 쇼트리드의 개수가 너무 많아지면 그 과정에서 시간이 너무 많이 소요된다.